Group Members:
Yash Kambli - yash.kambli22@spit.ac.in
Ayush Nemade - ayush.nemade22@spit.ac.in

Comprehensive Summary: Build LLM Research Assignment

Study Overview and Methodology

This research investigates the process of training a Large Language Model (LLM) from scratch using classic English literature. It examines how architectural choices and hyperparameters influence performance. The study uses three novels—*Pride and Prejudice*, *Dracula*, and *A Room with a View*—with a 90:10 train-validation split. It systematically explores variations in learning rates, model depth, attention mechanisms, and training duration, measuring both quantitative metrics and qualitative output quality.

Learning Rate Optimization

The experiments identified 0.0001 as the optimal learning rate, offering the best balance between training stability and convergence speed. Higher learning rates (0.001 and 0.01) caused unstable training, with the highest rate leading to divergence and unusable outputs. A learning rate of 0.0005 showed good early convergence but began to overfit after 8 epochs. This highlights how critical learning rate tuning is for stable and generalizable LLM training.

Architecture Design Considerations

Increasing the number of transformer layers from 7 to 20 significantly improved training performance—reducing training loss from 1.5563 to 0.6342—and enhanced output coherence. However, deeper models showed diminishing returns on validation performance, suggesting overfitting on the limited dataset. For attention mechanisms, using 12 to 16 heads provided the best trade-off between performance and efficiency. Although increasing to 24 heads slightly improved results, the gains did not justify the added computation. The 12-head setup delivered the most balanced validation performance (loss: 4.4537).

Training Duration and Overfitting

Prolonged training exposed serious overfitting risks. While training loss dropped nearly to zero after 50 epochs (0.0232), validation loss increased to 7.4341. At 100 epochs, the problem worsened (training loss: 0.0030, validation loss: 9.1142). These results underline the importance of early stopping techniques to avoid memorization and maintain generalization, especially when using small datasets.

Critical Architectural Components

Ablation studies showed the importance of key transformer components. Removing normalization layers or residual connections led to unstable training (validation losses of 6.7815 and 6.8086, respectively) and poor-quality outputs. Omitting the feedforward network allowed

partial training (validation loss: 4.6727) but significantly weakened the model's abstraction capabilities. These results confirm the complementary and essential roles of normalization, residuals, and feedforward blocks in stable transformer training.

Practical Implications

This study shows that successful LLM training requires a well-balanced model architecture, appropriate regularization, and careful hyperparameter tuning. The most effective configuration in this study featured 20 transformer layers, 12 to 16 attention heads, a learning rate of 0.0001, and early stopping. The research also emphasizes the non-negotiable importance of normalization and residual connections for model stability. These findings offer practical insights for training resource-efficient language models on limited data.